Trajectoires d'emploi identifiées à l'aide de cartes auto-organisées classifiantes. Etude réalisée avec le Panel Study of Income Dynamics, 1993-2003

Come Etienne[1], Cottrell Marie[1] and Gaubert Patrice[2] *

1- SAMM - Université Paris 1 Panthéon-Sorbonne
90, rue de Tolbiac, Paris 13 - France

2- ERUDITE - Université Paris 12
60, avenue du Général de Gaulle, 94100 Créteil - France

Abstract. Nous étudions les situations de chefs de famille américains vis-à-vis de l'emploi. Nous proposons pour cela une méthode de classification et de visualisation de données de grande dimension basée sur l'algorithme de Kohonen permettant l'apprentissage des cartes auto-organisatrices. Cette méthode que nous nommerons cartes auto-organisée classifiante nous permet d'étudier la segmentation du marché du travail américain en classes aux caractéristiques différentes et d'étudier la dynamique des trajectoires professionnelles des foyers appartenant au panel parmi ces classes.

1 Présentation du problème

La notion de trajectoire correspond à l'observation de changement de situation sur le marché, les situations possibles étant identifiées à partir des variables mesurant les principales caractéristiques de l'emploi occupé ou des différents types d'interruption d'activité entre 2 emplois. A la base de ce travail se trouve la théorie du marché segmenté (Cf. Doeringer [1]) qui, essentiellement, distingue un segment primaire (bons emplois avec grande stabilité, rémunérations supérieures à la moyenne, progression de responsabilités et rémunérations) et un segment secondaire (emplois en général peu qualifiés, peu stables ou à durée limitée, à temps partiel, à rémunération inférieure à la moyenne et peu de progression dans le temps) en dehors du chômage proprement dit.

L'utilisation d'un panel (Panel Study Of Income Dynamics) [1] permet d'obtenir la situation de plusieurs milliers de salariés pendant chacune des années de la période étudiée et d'obtenir ainsi leur trajectoire au travers de cette segmentation. Une première étude [2] portant sur la période 1984-1992, a permis d'obtenir une première segmentation. La majeure partie des salariés (autour des 2/3, selon la classe occupée l'année de départ) ne change pas de situation, ce qui indique globalement une assez grande stabilité pour cette première période caractérisée

*e-mail adresses : come@univ-paris1.fr, cottrell@univ-paris1.fr, patrice.gaubert@univ-paris12.fr

[1] Voir sur le site http://psidonline.isr.umich.edu les données et les références bibliographiques

par un chômage important qui tend à se réduire dans l'économie américaine au tournant des années 90.

Une analyse des transitions observées doit permettre de répondre à des questions essentielles pour la compréhension du marché, en particulier pour la définition de politiques de gestion active du marché du travail, par exemple :

- quelle transition s'opère pour les individus qui connaissent un choc de chômage en étant dans le segment primaire plutôt que dans le secondaire (quelle suite, quel temps d'attente) ?
- la trajectoire chômage $\rightarrow$ emploi précaire $\rightarrow$ "bon emploi" a t-elle une réalité et à quelles conditions ?
- y a t-il un lien entre la durabilité de l'emploi primaire et la trajectoire suivie antérieurement ?

Ces interrogations sont sans doute encore plus cruciales dans la période suivante, 1993-2003, compte tenu des transformations majeures qu'a connues le marché américain avec le développement rapide de nouvelle formes précaires d'emploi et la réduction globale de la stabilité (cf. par exemple Neumark [3]).

Il s'agit donc dans cette étude :

- de reprendre, sur la période suivante, la mise en évidence de situations sur le marché du travail bien différenciées, interprétables en termes de segments ;
- de caractériser ces classes à l'aide de variables qualitatives (niveau d'études, âge, sexe") ;
- sur la base du schéma théorique et avec l'a priori des constatations faites sur la période précédente, le nombre de macro-classes choisi est de 5 (quelques essais insatisfaisants ont été faits avec 6 et 7 macro-classes).

L'analyse des transitions constatées entre grands types de situations est toujours au centre des préoccupations. Pour déterminer ces classes et ces transitions, nous proposons un algorithme qui s'inspire de l'algorithme de Kohonen ([4]).

2 Cartes auto-organisées classifiantes

Les cartes auto-organisées sont des outils d'analyse de données bien connus. Elles sont généralement utilisées pour projeter des données de grande dimension sur un espace discret de faible dimension (généralement uni ou bi-dimensionnel). Elles offrent donc des potentialités intéressantes en terme de visualisation et d'exploration des données, en plus de leur intérêt intrinsèque en classification.

L'algorithme de Kohonen utilisé pour construire une carte auto-organisée est un algorithme de classification respectant la topologie de l'espace des observations. Comme la plupart des algorithmes de classification, il regroupe les observations en un certain nombre de classes K, et construit un ensemble de

vecteurs appelés vecteurs-codes notés $\mathbf{m}_i$, $i \in \{1, \ldots, K\}$ représentant chacun une classe. Les vecteurs-codes étant construits, les observations sont affectées à la classe dont le vecteur-code est le plus proche (au sens d'une distance donnée). Cet algorithme se différencie des algorithmes de classification classiques par l'introduction d'une structure de voisinage a priori entre les classes. Si l'on choisit une structure de voisinage telle que les classes soient disposées sur une grille plane en général carrée (ou sur une ficelle), l'algorithme possède des propriétés de visualisation très utiles pour représenter en deux dimensions (ou en une dimension) des données multidimensionnelles.

Chaque classe (sous-ensemble de l'espace des observations) est décrite par deux attributs :

- une position sur la carte c'est-à-dire un indice i sur une grille ;
- un vecteur-code dans l'espace des observations noté $\mathbf{m}_i$.

Construire une carte, c'est donc se donner une topologie entre les classes et construire un ensemble de vecteurs-codes. L'algorithme de Kohonen permet de trouver ces vecteurs codes, une fois la topologie entre les classes définie. Cet algorithme est dans sa forme classique présenté comme un algorithme itératif stochastique, défini de la manière suivante :

1. les vecteurs codes sont initialisés aléatoirement dans l'espace des observations ;
2. à chaque étape t, on modifie les vecteurs-codes $\mathbf{m}_i(t)$ de la manière suivante :
   - on tire aléatoirement une observation $\mathbf{x}_{t+1}$ et on réalise deux étapes;
   - *Compétition,* on détermine la classe gagnante (parmi toutes les classes) pour l'observation $\mathbf{x}_{t+1}$
   - *Coopération,* on modifie les vecteurs codes de la classe gagnante et de ses voisines sur la carte (afin de les rapprocher de l'observation $\mathbf{x}_{t+1}$)

Dans la pratique, l'algorithme est stoppé lorsque les vecteurs codes ne bougent plus beaucoup ou lorsqu'un nombre maximal d'itérations a été effectué. Différentes fonctions de voisinage sont classiquement utilisées telles que :

$$h(t, c, i) = \mathbf{1}_{(d(c,i)<\sigma_t)} \quad (1)$$

$$h(t, c, i) = \exp(-d(c, i)^2/2\sigma_t^2). \quad (2)$$

où $d(c, i)$ est la distance sur la grille entre la classe c et la classe i.

Il est aisé d'utiliser d'autres structures de voisinage entre classes que celle qui est définie à l'aide d'une grille rectangulaire ou d'une ficelle. En effet, la structure de voisinage intervient dans cet algorithme au travers de $d(c, i)$ qui est la distance sur la grille entre la classe c et la classe i. On peut modifier cette distance, et ainsi définir une structure de voisinage différente, qui peut se révéler plus pertinente qu'une grille ou qu'une ficelle pour certains jeux de données.

La théorie des graphes permet de définir de telles distances aisément comme cela a déjà été noté par plusieurs auteurs [5, 6]. Il suffit pour cela de définir un graphe entre les classes (chaque classe correspondant alors à un sommet du graphe) et d'utiliser la distance du plus court chemin entre les sommets du graphe, définie par le nombre minimum d'arêtes devant être parcourues pour rejoindre un sommet en partant de l'autre. Nous proposons donc de modifier l'algorithme en prenant en entrée une matrice d'adjacence qui spécifie le graphe voulu par l'utilisateur. Tous les graphes (non-orientés) peuvent théoriquement être utilisés, cependant un type de graphe semble plus intéressant : les graphes planaires. De tels graphes peuvent en effet être représentés en deux dimensions ce qui permet de conserver les avantages des cartes auto-organisées en grille en terme de visualisation et d'exploration des données.

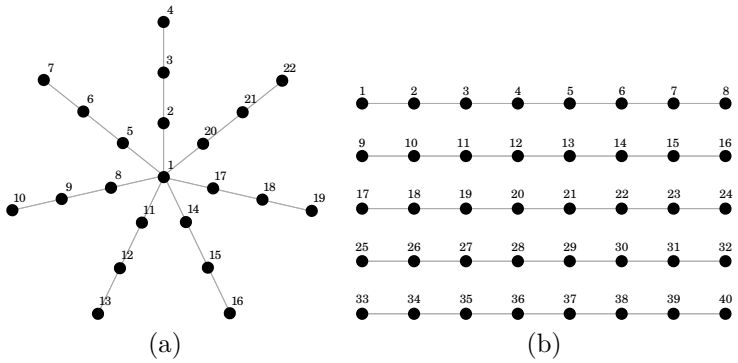


Fig. 1: (a) représentation planaire d'un graphe en étoile avec 7 branches de 3 unités; (b) représentation planaire d'un graphe non-connexe constitué de 5 ficelles de 8 unités.

Certaines topologies présentent en outre un intérêt particulier pour utiliser des informations a priori sur la structure du jeu de données étudié. Une structure telle que celle présentée sur la figure 1 (a) peut par exemple être intéressante dans le cadre du diagnostic de système industriel où l'on recherche des déviations par rapport à un état de bon fonctionnement [7]. Ici nous nous intéressons aux topologies définies à l'aide de graphes non connexes, ce qui est adapté au cas où on recherche un nombre connu de macro-classes.

En effet, lorsque le graphe utilisé pour définir la topologie d'une carte n'est pas connexe, l'étape de "coopération" de l'algorithme ne concerne que les unités appartenant à la même composante du graphe que l'unité gagnante. L'étape de compétition de l'algorithme n'étant pas modifiée, l'algorithme obtenu atteindra un double objectif :

1. classer les observations dans des macro-classes correspondant aux différentes composantes connexes du graphe ;

2. organiser les classes à l'intérieur des macro-classes.

Les vecteurs-codes sont alors doublement indicés m_{ij}, $i \in \{1, \ldots, K\}$, $j \in \{1, \ldots, n_i\}$. Si l'on note $d((i,j),(i',j'))$ la distance du plus court chemin (dans le graphe) entre les classes (i,j) et (i',j'), celle-ci est égale à $+\infty$ et $h(t,(i,j),(i',j')) = 0$ si $i \neq i'$. Les vecteurs codes des classes n'appartenant pas à la même macro-classe que l'unité gagnante ne sont pas affectés par l'étape de coopération. L'algorithme peut donc être ré-écrit de la manière suivante :

1. les vecteurs codes sont initialisés aléatoirement dans l'espace des observations ;
2. à chaque étape t, on modifie les vecteurs-codes $\mathbf{m}_{ij}(t)$ de la manière suivante :
   - on tire aléatoirement une observation $\mathbf{x}_{t+1}$ et on réalise deux étapes;
   - *Compétition,* on détermine la classe gagnante (parmi toutes les classes) pour l'observation $\mathbf{x}_{t+1}$ par l'équation :

$$[i^*(t+1), j^*(t+1)) = \arg \min_{i \in \{1,\ldots,K\}, j \in \{1,\ldots,n_i\}} ||\mathbf{x}_{t+1} - \mathbf{m}_{ij}(t)||; \quad (3)$$

   - *Coopération,* on modifie les vecteurs codes de la classe gagnante et de ses voisines (qui appartiennent forcément à la même macro-classe i^*) par :

$$\mathbf{m}_{i^*j}(t+1) = \mathbf{m}_{i^*j}(t) + \alpha(t) h(t,(i^*,j^*),(i^*,j)) \left[\mathbf{x}_{t+1} - \mathbf{m}_{i^*j}(t)\right], \quad (4)$$

   où t est le numéro de l'itération, $\alpha(t)$ le paramètre d'apprentissage de l'algorithme et $h(t,(i^*,j^*),(i^*,j))$ la fonction de voisinage à l'étape t entre les classes (i^*,j^*) et (i^*,j).

En conclusion, en limitant la coopération qui ne se fait plus qu'à l'intérieur des macro-classes et en conservant une compétition entre toutes les classes, cet algorithme permet d'obtenir (comme nous allons le voir dans l'exemple étudié ici et comme illustré par la figure 2), une classification en un nombre fixé à l'avance de macro-classes elles-mêmes auto-organisées.

Ce type de topologie est pertinent dans le contexte de la segmentation du marché du travail, puisque l'on recherche comme indiqué dans l'introduction, une segmentation en 5 macro-classes faciles à décrire et elles-mêmes segmentées en classes organisées. Plusieurs questions mériteraient d'être étudiées plus en profondeur dans le cas général, en particulier la question du choix du nombre de macro-classes qui peut devenir cruciale si aucune information n'est disponible a priori.

Nous allons maintenant décrire et présenter les résultats obtenus avec une telle méthodologie.

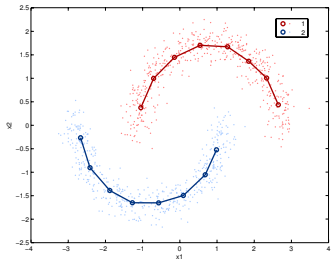


Fig. 2: Exemple de carte de Kohonen classifiante sur des données simulées dans $\mathbb{R}^2$. La figure représente un nuage de points simulés, la position des vecteurs codes des deux macro-classes et le graphe utilisé pour définir la topologie de la carte (2 macro-classes de 8 unités).

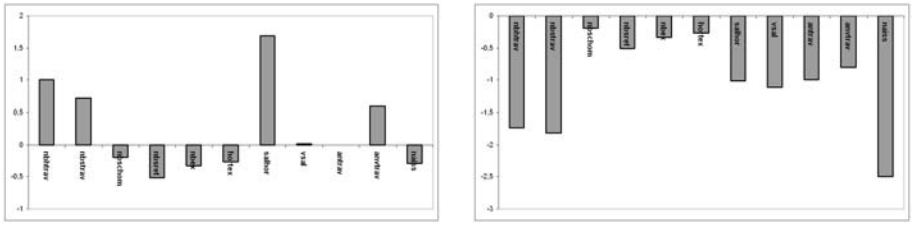
3 Les données

Nous disposons de 3513 ménages observés tous les deux ans de 1993 à 2003, les années impaires, et donc de 6 observations par ménage. La situation du ménage est décrite par la situation du chef de ménage, essentiellement des hommes. Le nombre de variables renseignées dans l'enquête est très important et beaucoup sont redondantes. Nous en avons choisi 15, les plus pertinentes pour notre travail. Elles se répartissent en deux groupes : 10 variables décrivant la situation sur le marché du travail, 5 variables caractéristiques du chef de famille. Les 10 premières variables sont des variables réelles utilisées pour classer les ménages grâce à une carte de Kohonen classifiante. Les cinq variables suivantes sont binaires (sexeh, prop) ou discrètes (les trois autres). Elles ne sont pas utilisées dans les classifications, mais permettent une description des classes obtenues. Les significations des différentes variables sont listées ci dessous.

Les variables décrivant la situation des individus dans le marché de l'emploi pour chaque année sont : nombre d'heures travaillées par semaine, (*nbhtrav*), nombre de semaines travaillées dans l'année, (*nbstrav*), nombre de semaines chômées, (*nbschom*), nombre de semaines en retrait du marché du travail, (*nbsret*), nombre d'emplois secondaires, (*nbex*), nombre d'heures en emplois secondaires sur l'année, (*hortex*), salaire horaire, (*salhor*), ancienneté sur le marché du travail, (*antrav*), année de naissance, (*naiss*).

Les variables liées aux individus sont sexe, (*sexeh*), nombre d'enfants, (*nbenf*), niveau d'études, (*etudeh*), taille du ménage, (*taille*), propriété du logement, (*prop*).

Pour effectuer les classifications et construire les cartes de Kohonen, nous considérons les observations comme des couples (une année, un ménage décrit par les 10 variables mesurées cette année-là), soit un couple (i, j), où $i = 1993, 1995, \ldots, 2003$, et $j = 1, 2, \ldots, 3513$. La figure 3 montre deux exemples d'individus, en 1995, les données étant centrées et réduites.



(a) (b)

Fig. 3: Deux individus observés en 2003, (a) temps de travail important, salaire élevé,... (b) temps de travail inférieur à la moyenne, salaire faible, ...

4 Résultats

On considère une carte de Kohonen classifiante composée de 5 macro-classes (non numérotées) comprenant 8 unités chacune. Les 10 variables sont tout d'abord centrées et réduites pour limiter les effets d'échelle. La classification obtenue peut être visualisée et étudiée sous différents angles pour identifier les macro-classes et de donner à celles-ci une sémantique économique.

4.1 Définition et identification des macro-classes

Tout d'abord, nous représentons, pour chacune des classes, les valeurs des composantes du vecteur code correspondant, sous forme de barres. La figure 4 présente les résultats obtenus ; les différentes macro-classes sont pour cela disposées en ligne. Comme il n'y a pas de relation d'ordre entre les macro-classes, elles ont été renumérotées a posteriori de 1 à 5 après analyse de leur contenu.

Nous pouvons observer que les macro-classes possèdent des profils bien distincts :

C1 individus qui se retirent du marché du travail parce qu'ils sont âgés en fin d'activité, ou pour des raisons personnelles,

C2 individus en grande précarité : à la gauche de la ficelle C2, précarité et activité partielle, à droite chômage permanent,

C3 emplois du segment secondaire, en moyenne 41 heures par semaine 48 semaines dans l'année pour un salaire faible et une faible ancienneté dans l'emploi,

C4 individus exerçant au moins deux activités simultanément,

C5 emplois du segment primaire, salaire supérieur à la moyenne, un seul emploi à temps complet toute l'année, ancienneté dans l'emploi nettement supérieure à la moyenne.

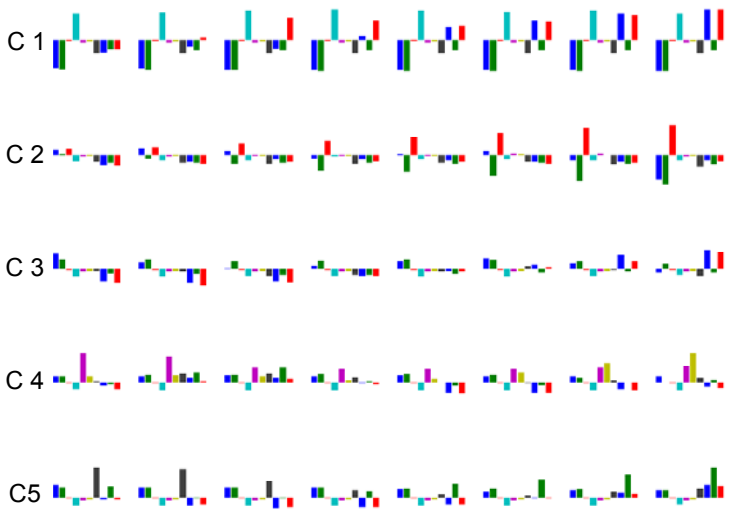


Fig. 4: Représentation sous forme de barres des vecteurs-codes de la carte de Kohonen classifiante obtenue sur le panel étudié. Les 10 composantes des vecteurs sont dans l'ordre : nbhtrav, nbstrav, nbschom, nbsret, nbex, hortex, salhor,antrav,anctrav,naiss.

L'étude de la répartition des 5 variables qualitatives dans les 5 macro-classes permet de compléter la description. Par exemple la classe C1 comprend une proportion de femmes très supérieure à sa valeur moyenne dans le panel ; les individus de la classe 5 ont un niveau d'études supérieur à la moyenne, etc.

On peut aussi compléter cette description en représentant toutes les variables pour les différentes unités de chaque macro-classe, voir la figure 5 qui montre bien l'organisation interne à chaque macro-classe mise en place par l'algorithme de manière automatique. Pour toutes les macro-classes, sauf la macro-classe C3, une ou deux variables permettent de comprendre l'organisation interne de la macro-classe. Par exemple la macro-classe C2 est organisée des situations les moins précaires à gauche aux situations les plus précaires à droite comme le montre l'évolution de la variable ($nbschom$)§. En ce qui concerne C5 les rémunérations horaires sont croissantes de droite à gauche alors que l'ancienneté dans l'emploi suit le mouvement inverse.

4.2 Transitions entre classes

On définit ensuite la trajectoire d'un individu par la succession des numéros des classes $(i, j), i = 1, \ldots, 5, j = 1, \ldots, 8$ auxquelles il appartient aux 6 dates retenues (1993, 1995, ..., 2003). On représente dans la figure 6 trois exemples de telles trajectoires.Pour étudier les transitions, on se borne à considérer les transitions entre les 5 macro-classes. Par simple comptage, on calcule les prob-

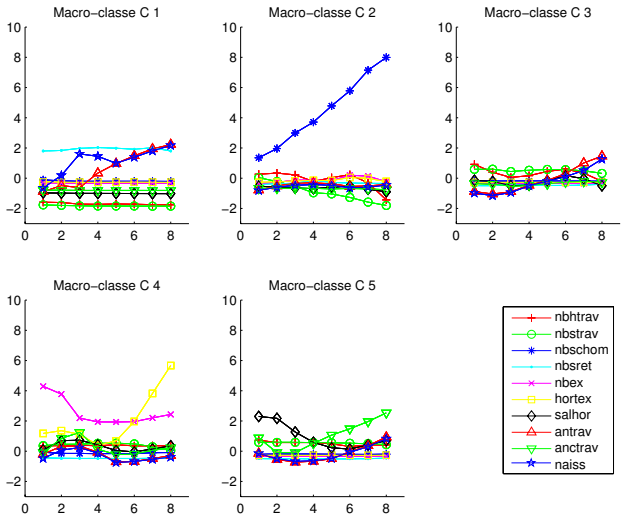
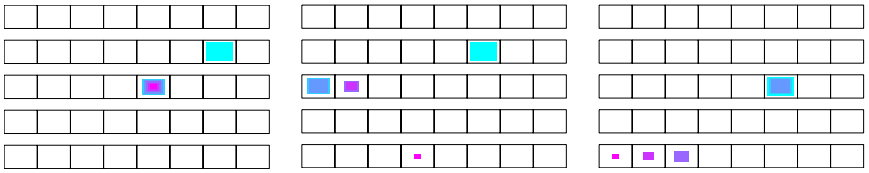


Fig. 5: Représentation de l'organisation interne des classes : les abscisses correspondent aux numéros de l'unité à l'intérieur de la classe considérée, les ordonnées aux valeurs des vecteurs codes correspondants.

abilités empiriques d'appartenir à la classe j l'année n, sachant qu'on est dans la classe i l'année précédente.



(a) (b) (c)

Fig. 6: Exemples de trajectoires individuelles sur la carte. Le dégradé de couleur ainsi que la taille du carré permet de représenter l'année de l'enquête, les carrés de grande taille et bleu clair correspondent à la situation du chef de famille en 1993; les petits carrés roses à sa situation en 2003.

La matrice obtenue est une matrice de Markov. Les éléments de la diagonale (probabilités de rester dans la même macro-classe sont très fortes pour les classes, C1, C3, C5, importante (48 %) pour la classe C4 dont les individus exercent plusieurs emplois, et beaucoup plus faible (21 %) pour la classe C2, celle des emplois précaires et des chômeurs. La sortie de la classe C2 se fait essentiellement vers la classe C3 (segment secondaire). Pour un individu qui quitte la classe C4 (deux emplois ou plus), il a trois chances sur cinq de passer au segment secondaire et deux chances sur cinq de passer au segment primaire.

0.89	0.03	0.07	0.01	0.01
0.14	0.21	0.49	0.08	0.08
0.07	0.05	0.64	0.09	0.16
0.02	0.02	0.28	0.48	0.20
0.02	0.02	0.13	0.06	0.77

Table 1: Matrice de transition entre les classes

	C1	C2	C3	C4	C5
Loi stationnaire calculée	0.2832	0.0345	0.2818	0.0970	0.3035
Distribution moyenne observée	0.20	0.04	0.33	0.11	0.32

Table 2: comparaison de la répartition calculée (sous l'hypothèse de stationnarité) et de la répartition empirique

Si on faisait l'hypothèse que la situation globale (économique et règlementaire) n'a pas changé au cours de la période 1993-2003, on pourrait calculer la loi stationnaire correspondante (comme limite des lignes des puissances successives de la matrice de transition), et comparer les proportions calculées avec les proportions observées, voir table 2.

En pratiquant un test du Chi-deux d'ajustement, on conclut que ces deux lois ne sont pas identiques, et il resterait à étudier les lois stationnaires estimées année après année.

5 Perspectives

Une petite partie de l'échantillon analysé dans la 1ère étude (années 1984-1992, voir [2]) se retrouve dans celui de la 2ème : avec précaution, du fait qu'il n'y a qu'environ 400 ménages concernés, il sera possible d'observer l'évolution des types de transition pour les mêmes individus alors qu'ils se trouvent dans deux modes de fonctionnement du marché assez différents.

Plus globalement, nous pensons comparer la situation du marché du travail au cours des deux périodes étudiées, et identifier les mécanismes économiques en jeu.

D'autre part, une dimension importante de la question étudiée ne peut pas être prise en compte sérieusement dans l'état actuel de la table de données : l'importance du genre des personnes. Ceci résulte de la convention statistique usuelle qui s'appuie sur la notion de chef de famille pour rassembler les informations, ce chef étant systématiquement l'homme du ménage s'il y en a un ; les données collectées ici étant celles des chefs de famille présents dans la période, les femmes ne représentent que 25% de l'échantillon et dans une situation de chef d'une famille monoparentale, ce qui n'est pas indifférent par rapport aux décisions prises sur le marché du travail. La récupération des données sur les con-

joints des chefs de famille (données personnelles et données du marché du travail) permettra de reprendre la classification sur des individus hommes et femmes en proportions voisines, certains disposant d'un conjoint et d'autres pas. On sait que ceci intervient dans les décisions de sortie, de choix d'emploi, de possibilité de refus d'un emploi précaire pour trouver un emploi durable, etc.

Enfin, sur des trajectoires significatives statistiquement, on mettra en évidence leurs facteurs déterminants à l'aide de modèles de régression à variables dépendantes qualitatives.

References

[1] P. B. Doeringer and M. J. Piore. *Internal Labor Market and Manpower Analysis.* D.C. Heath and Company, Lexington Massachussetts, 1971.

[2] P. Gaubert and M. Cottrell. A dynamic analysis of segmented labor market. *Fuzzy Economic Review*, IV:62–82, 1999.

[3] D. Neumark, D. Polsky, and D. Hansen. Has job stability declined yet ? new evidence for the 90's. *Journal of Labour Economics*, 17, 1999.

[4] T. Kohonen. *Self-Organizing Maps.* Information siences. Springer, 1995.

[5] A. Barsi. Neural self-organization using graphs. In *Proceedings of the Third International Conference Machine Learning and Data Mining in Pattern Recognition, Leipzing (Germany)*, volume 2734 of *Lecture Notes in Artificial Intelligence*, pages 343–352. Springer, July 2003.

[6] J. Pakkanen, J. Iivarinen, and E. Oja. The elvoving tree - analysis and applications. *IEEE Transactions on Neurals Networks*, 17:591–603, 2006.

[7] E. Côme, M. Cottrell, M. Verleyssen, and J. Lacaille. Self organizing star (sos) for health monitoring. In *Proceedings of the European conference on artificial neural networks, Bruges (Belgium)*, pages X–X, 2010.